

# Aprendizaje Automático y Minería de Datos

Presentación de la asignatura

Cristina Tîrnăucă

Dept. Matesco, Universidad de Cantabria

Fac. Ciencias – Grado en Ing. Informática

# Cuestiones Factuales

De índole práctica

## Personal e infraestructura

- ▶ Clases a cargo de Cristina Tîrnăucă (cristina.tirnauca@unican.es).

# Cuestiones Factuales

De índole práctica

## Personal e infraestructura

- ▶ Clases a cargo de Cristina Tîrnăucă (cristina.tirnauca@unican.es).
- ▶ Horario:
  - ▶ desarrollo teórico, ejemplos y ejercicios, en general los martes y jueves (11:45 - 12:45) en el **Seminario de Informática**,
  - ▶ laboratorio el lunes (11:45 - 13:45) en el **LSC 2**:
    - ▶ sesiones de prácticas **puntuables** en python hasta Semana Santa, aproximadamente, y
    - ▶ prácticas **no puntuables** en weka y knime hasta el final del cuatrimestre.

# Cuestiones Factuales

De índole práctica

## Personal e infraestructura

- ▶ Clases a cargo de Cristina Tîrnăucă (cristina.tirnauca@unican.es).
- ▶ Horario:
  - ▶ desarrollo teórico, ejemplos y ejercicios, en general los martes y jueves (11:45 - 12:45) en el **Seminario de Informática**,
  - ▶ laboratorio el lunes (11:45 - 13:45) en el **LSC 2**:
    - ▶ sesiones de prácticas **puntuables** en python hasta Semana Santa, aproximadamente, y
    - ▶ prácticas **no puntuables** en weka y knime hasta el final del cuatrimestre.
- ▶ Información actualizada sobre el desarrollo de la asignatura en: `moodle.unican.es`

# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo.

# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo. Se permiten retrasos en la entrega de prácticas de máximo dos semanas:
  - ▶ si el retraso no supera una semana, la nota baja en 1 punto sobre 10;
  - ▶ para retrasos de más de una semana, la nota baja en 3 puntos sobre 10.

# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo. Se permiten retrasos en la entrega de prácticas de máximo dos semanas:
  - ▶ si el retraso no supera una semana, la nota baja en 1 punto sobre 10;
  - ▶ para retrasos de más de una semana, la nota baja en 3 puntos sobre 10.
- ▶ Práctica **individual** (hasta 2 puntos): a partir de un “dataset” que acordemos, harás entrar en juego todo lo que hayas aprendido e intentarás completar un miniproyecto de Minería de Datos (el ingrediente básico es la iniciativa personal).

# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo. Se permiten retrasos en la entrega de prácticas de máximo dos semanas:
  - ▶ si el retraso no supera una semana, la nota baja en 1 punto sobre 10;
  - ▶ para retrasos de más de una semana, la nota baja en 3 puntos sobre 10.
- ▶ Práctica **individual** (hasta 2 puntos): a partir de un “dataset” que acordemos, harás entrar en juego todo lo que hayas aprendido e intentarás completar un miniproyecto de Minería de Datos (el ingrediente básico es la iniciativa personal).
- ▶ Cuestiones y problemas puntuales en moodle (hasta 1 punto)



# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo. Se permiten retrasos en la entrega de prácticas de máximo dos semanas:
  - ▶ si el retraso no supera una semana, la nota baja en 1 punto sobre 10;
  - ▶ para retrasos de más de una semana, la nota baja en 3 puntos sobre 10.
- ▶ Práctica **individual** (hasta 2 puntos): a partir de un “dataset” que acordemos, harás entrar en juego todo lo que hayas aprendido e intentarás completar un miniproyecto de Minería de Datos (el ingrediente básico es la iniciativa personal).
- ▶ Cuestiones y problemas puntuales en moodle (hasta 1 punto)

La calificación obtenida en el examen final (de 0 a 10 puntos) se multiplica por 0,5 y se suma a la nota de curso **sólo si es  $\geq 4$** .

# Evaluación

A lo largo del curso se obtiene una **nota de curso** en  $[0, 5]$ .

- ▶ Prácticas **en grupo** de 2 o 3 personas (hasta 2 puntos): la calificación es la misma para todos los miembros del grupo. Se permiten retrasos en la entrega de prácticas de máximo dos semanas:
  - ▶ si el retraso no supera una semana, la nota baja en 1 punto sobre 10;
  - ▶ para retrasos de más de una semana, la nota baja en 3 puntos sobre 10.
- ▶ Práctica **individual** (hasta 2 puntos): a partir de un “dataset” que acordemos, harás entrar en juego todo lo que hayas aprendido e intentarás completar un miniproyecto de Minería de Datos (el ingrediente básico es la iniciativa personal).
- ▶ Cuestiones y problemas puntuales en moodle (hasta 1 punto)

La calificación obtenida en el examen final (de 0 a 10 puntos) se multiplica por 0,5 y se suma a la nota de curso **sólo si es  $\geq 4$** .

**Importante:** La nota de la evaluación continua (problemas, prácticas en grupo y práctica individual) se tiene en cuenta sólo para el examen del periodo ordinario. En el periodo de recuperación, el examen tiene un peso de 100%.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.
  - ▶ Por ejemplo, en **datos** disponibles.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.
  - ▶ Por ejemplo, en **datos** disponibles.
  - ▶ Pero si tenemos todos los datos, no hay nada a predecir.

# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.
  - ▶ Por ejemplo, en **datos** disponibles.
  - ▶ Pero si tenemos todos los datos, no hay nada a predecir.
- ▶ Ingrediente imprescindible: la **incertidumbre**.



# Análisis de Datos

## Construcción de modelos descriptivos o predictivos

### Objetivo:

Una ventaja económica o (menos frecuentemente) humana.

- ▶ La intención es lograrla mediante **predicciones acertadas**, al menos parcialmente.
- ▶ Predecir al azar difícilmente proporciona ventajas: queremos hacerlo mejor que al azar.
  - ▶ Para ello, habremos de basarnos en algo.
  - ▶ Por ejemplo, en **datos** disponibles.
  - ▶ Pero si tenemos todos los datos, no hay nada a predecir.
- ▶ Ingrediente imprescindible: la **incertidumbre**.
- ▶ De las muchas maneras de gestionar el conocimiento incierto, la más relevante en data mining (que no la única) es el enfoque **estadístico**, basado en la **teoría de la probabilidad**.

# Minería de Datos

## Interés en realidades existentes

El proceso de minería de datos **incluirlá** fases de **modelado** a partir de observaciones (datos) sobre una **realidad compleja y existente**.

Taxonomía:

- ▶ Modelos descriptivos:
  - ▶ Segmentación
  - ▶ Asociación
- ▶ Modelos predictivos:
  - ▶ Regresión
  - ▶ Clasificación
- ▶ Modelos supervisados
- ▶ Modelos no supervisados

(Nociones mutuamente no excluyentes.)

# Minería de datos y el aprendizaje automático

## *Data Mining vs Machine Learning*

Estos dos términos son muchas veces confundidos, ya que a menudo emplean los mismos métodos y se superponen de manera significativa.

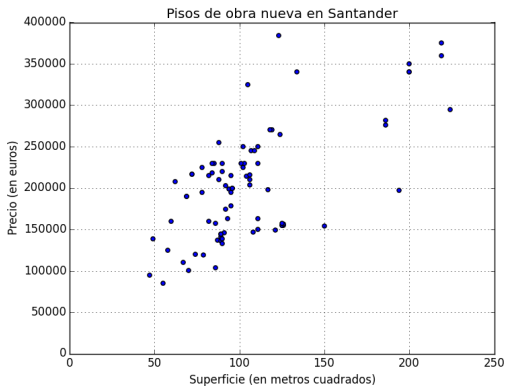
- ▶ el aprendizaje automático se centra en la predicción, basándose en propiedades **conocidas** extraídas de los datos de entrenamiento,
- ▶ la minería de datos se centra en el descubrimiento de propiedades (antes) **desconocidas** en los datos.

Las dos áreas se superponen en muchos sentidos:

- ▶ la minería de datos utiliza muchos métodos de aprendizaje automático, pero a menudo con un objetivo ligeramente diferente en mente,
- ▶ el aprendizaje automático también cuenta con métodos de minería de datos como por ejemplo “el aprendizaje no supervisado” como un paso de procesamiento previo para mejorar la precisión del modelo.

# Ejemplos

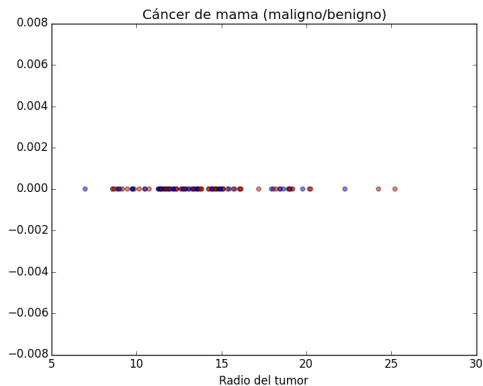
## Regresión



Otras variables: número de habitaciones, número de baños, si tiene ascensor, calefacción, trastero, parking, si la comunidad tiene piscina, si el piso está situado en el centro, ...

# Ejemplos

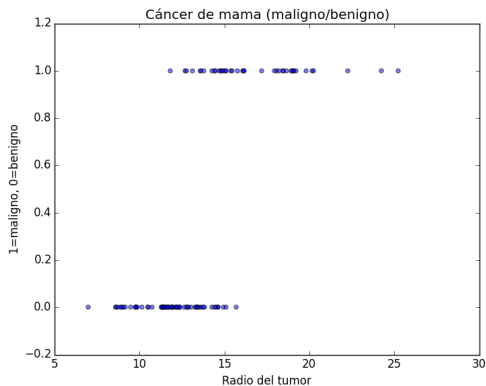
## Clasificación



Otras observaciones clínicas: la edad del paciente, el espesor del tumor, la homogeneidad del tamaño celular, la homogeneidad de la forma celular,...

# Ejemplos

## Clasificación



Otras observaciones clínicas: la edad del paciente, el espesor del tumor, la homogeneidad del tamaño celular, la homogeneidad de la forma celular,...

# Ejemplos

## Segmentación

Google News


← → ↻ https://news.google.com

Google

U.S. edition Modern

News

Top Stories



Washington Post

See realtime coverage

### Trump warns Israel that new settlements 'may not help' achieve Middle East peace

Washington Post - 5 hours ago

The White House on Thursday gently warned Israel that new or expanded settlements in the West Bank 'may not be helpful' in achieving a Middle East peace, while insisting it has no 'official position on settlement activity.'










Trump to Israel: Settlements 'may not help' achieve peace in the Middle East Fox News


Israeli Settlements 'May Not Be Helpful' for Middle East Peace, Trump Administration Says Wall Street Journal

From Israel: Settlements and 'The Ultimate Deal': Trump's Surprising Statement on Israel in Context Haaretz

Trending: Israel to build entirely new settlement in West Bank CNN

Opinion: The settlers' victory Jerusalem Post Israel News






New York Times

### Uber CEO to Leave Trump Advisory Council After Criticism

New York Times - 5 hours ago

Travis Kalanick, the chief executive of Uber, in December 2014. He quit President Trump's economic advisory council on Thursday.




Yahoo News

### N. Korea nuclear attack would trigger 'overwhelming response': Mattis

Yahoo News - 1 hour ago

Any nuclear attack by North Korea would trigger an "effective and overwhelming" response, US Defence Secretary James Mattis said Friday as he sought to reassure Asian allies rattled by President Donald Trump's isolationist rhetoric.



New York Times

### What Snap's IPO Filing Reveals About the Company

New York Times - 8 hours ago

Snap, the parent of Snapchat, disclosed several important aspects of its business in its initial public offering document. The complete filing is here.

Sign in to get news on topics you care about. Learn more





Recent

Japan to push back if Trump meddles with BOJ independence: sources Reuters - 17 minutes ago

Turkey to meet Syrian opposition, rebel groups in Ankara on Friday: sources Reuters - 31 minutes ago







US reversal on transparency could sting Canadian, European oil companies Reuters - 1 hour ago

Weather for Cantabria, Spain

Today	Sat	Sun	Mon
			
15° 11°	14° 11°	11° 8°	17° 12°

The Weather Channel - Weather Underground - AccuWeather

Sports scores

	Today	Yesterday
NHL		
 BOS	3-5 Final	WSH 
 MIN	1-5 Final	CGY 
 COL	0-5 Final	LAK 

← → ↻

# Ejemplos

## Asociación

### En un censo estadounidenses:

- ▶ Husband  $\rightarrow$  Male, Married-civ-spouse
- ▶ Married-civ-spouse  $\rightarrow$  Husband, Male
- ▶ Not-in-family  $\rightarrow \leq 50K$
- ▶ Black  $\rightarrow \leq 50K$ , United-States
- ▶ Adm-clerical, Private  $\rightarrow \leq 50K$
- ▶ Self-emp-not-inc  $\rightarrow$  Male
- ▶  $\leq 50K$  , Sales  $\rightarrow$  Private
- ▶ hours-per-week:50  $\rightarrow$  Male
- ▶ Female, Some-college  $\rightarrow \leq 50K$
- ▶ Divorced  $\rightarrow \leq 50K$



# Objetivos

## Competencias específicas

- Entender los conceptos y la terminología de las técnicas de minería de datos.

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.
- ▶ Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones.

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.
- ▶ Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones.
- ▶ Elegir, para un problema concreto, qué técnicas de minería de datos son más apropiadas.

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.
- ▶ Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones.
- ▶ Elegir, para un problema concreto, qué técnicas de minería de datos son más apropiadas.
- ▶ Generar los modelos y patrones elegidos utilizando una herramienta o paquete de minería de datos.

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.
- ▶ Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones.
- ▶ Elegir, para un problema concreto, qué técnicas de minería de datos son más apropiadas.
- ▶ Generar los modelos y patrones elegidos utilizando una herramienta o paquete de minería de datos.
- ▶ Evaluar la calidad de un modelo, utilizando técnicas sencillas de evaluación (validación cruzada).

# Objetivos

## Competencias específicas

- ▶ Entender los conceptos y la terminología de las técnicas de minería de datos.
- ▶ Reconocer los beneficios del uso sistemático de técnicas de extracción de conocimiento para la obtención de modelos y patrones predictivos o descriptivos.
- ▶ Conocer las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones.
- ▶ Elegir, para un problema concreto, qué técnicas de minería de datos son más apropiadas.
- ▶ Generar los modelos y patrones elegidos utilizando una herramienta o paquete de minería de datos.
- ▶ Evaluar la calidad de un modelo, utilizando técnicas sencillas de evaluación (validación cruzada).
- ▶ Implementar un algoritmo de minería de datos específico.

# Bibliografía, I

1. Jiawei Han, Micheline Kamber, **Jian Pei**:

- ▶ **Data Mining: Concepts and Techniques**, Academic Press (2001), 2nd. Ed. Morgan Kaufmann Publishers (2006), **3rd. Ed. Elsevier (2012)**.

Pretende una orientación práctica.

2. David Hand, Heikki Mannila, Pádraic Smyth:

- ▶ **Principles of data mining**, MIT Press (2001)

Un “clásico”.

`ftp://gamma.sbin.org/pub/doc/books/Principles_of_Data_Mining.pdf`

3. Michael Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn:

- ▶ **Guide to Intelligent Data Analysis**, Springer (2010).

Basado en KNIME.

`http://www.2shared.com/document/1AKhLJ-4/Guide_to_Intelligent_Data_Anal.html`



# Bibliografía, II

## 4. Ian H. Witten, Eibe Frank, Mark A. Hall:

- ▶ **Data mining: Practical machine learning tools and techniques with Java implementations**, Elsevier (2000), 2nd. Ed. (2005), 3rd. Ed. Elsevier (2011)

Es el libro que acompaña a Weka.

<http://home.etf.rs/~vm/os/dmsw/Morgan.Kaufman.Publishers.Weka.2nd.Edition.2005.Elsevier.pdf>

## 5. Trevor Hastie, Robert Tibshirani, Jerome Friedman:

- ▶ **The elements of statistical learning: data mining, inference, and prediction**, 2nd. Ed. Springer (2009)

La base más estadística de la minería de datos.

## 6. G. James, D. Witten, Trevor Hastie, Robert Tibshirani:

- ▶ **An Introduction to Statistical Learning with Applications in R**, Springer (2013)

Según los autores, es la versión más ligera del libro anterior

<http://www.stanford.edu/~hastie/pub.htm>